GPT4RoI: Instruction Tuning Large Language Model on Region-of-Interest

Year: 2023 | *Citations: 303* | *Authors: Shilong Zhang, Pei Sun, Shoufa Chen*

Abstract

Visual instruction tuning large language model(LLM) on image-text pairs has achieved general-purpose vision-language abilities. However, the lack of region-text pairs limits their advancements to fine-grained multimodal understanding. In this paper, we propose spatial instruction tuning, which introduces the reference to the region-of-interest(RoI) in the instruction. Before sending to LLM, the reference is replaced by RoI features and interleaved with language embeddings as a sequence. Our model GPT4RoI, trained on 7 region-text pair datasets, brings an unprecedented interactive and conversational experience compared to previous image-level models. (1) Interaction beyond language: Users can interact with our model by both language and drawing bounding boxes to flexibly adjust the referring granularity. (2) Versatile multimodal abilities: A variety of attribute information within each RoI can be mined by GPT4RoI, e.g., color, shape, material, action, etc. Furthermore, it can reason about multiple RoIs based on common sense. On the Visual Commonsense Reasoning(VCR) dataset, GPT4RoI achieves a remarkable accuracy of 81.6%, surpassing all existing models by a significant margin (the second place is 75.6%) and almost reaching human-level performance of 85.0%. The code and model can be found at https://github.com/jshilong/GPT4RoI.